Data Mining Lab, Exercise 3

Team# 1: Poliakov Valerii, Holovnia Dmytro, Selvaraj Sinju

Datasets: f1.csv, zoo.csv

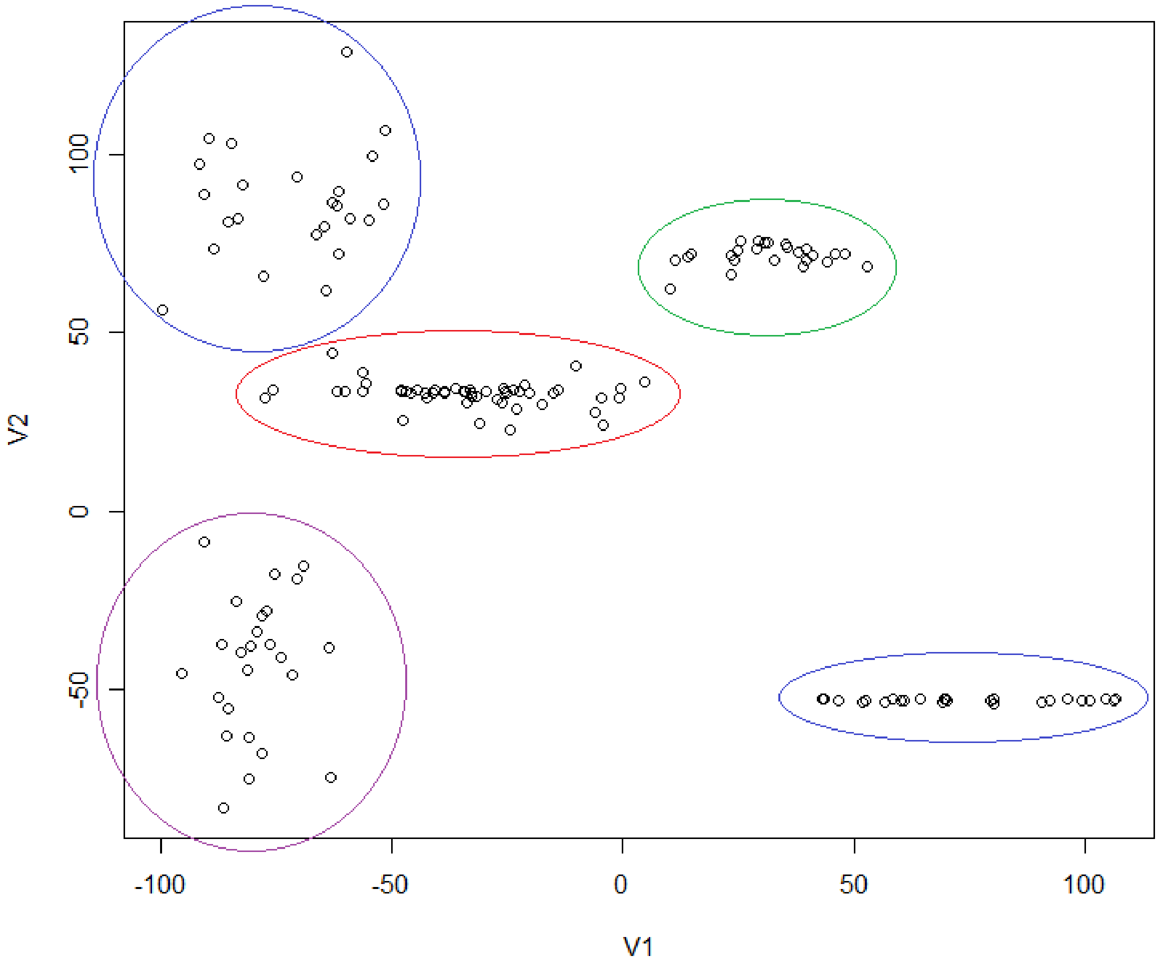
Task# 1

(3pt) Perform the calculations using the k-means algorithm for data from file "f#.csv".

a) Decide on the most likely number of clusters on the basis of the scatter plot of the data. For the chosen number of clusters, perform clustering using the k-means algorithm. Repeat the calculations 5 times. Write the values of SSB/TSS for each of the algorithm runs. Indicate the minimal and maximal value of SSB/TSS as well as its average value. Show the best solution (clusters) in the scatter plot. Indicate the centroids.

Step 1. Read the data. Draw plot and guess number of clusters.

```
d<-read.csv(file="f1.csv", header = FALSE, sep=' ')
d
with(d, plot(V1, V2))
```



Dataset in the f1.csv file has no headers and whitespace used as a separator. Plot shows five distinct groups. Most likely there are 5 clusters.

Step 2. Repeat 5 times k-means clustering for five clusters. Note the SBB/TSS values for each iteration. SBB (Sum of squares between clusters) and TSS (Total Sum of Squares) are used as measures to evaluate the performance of the algorithm.

Script command:

```
km<-kmeans(d, centers=5)
```

Iteration 1:

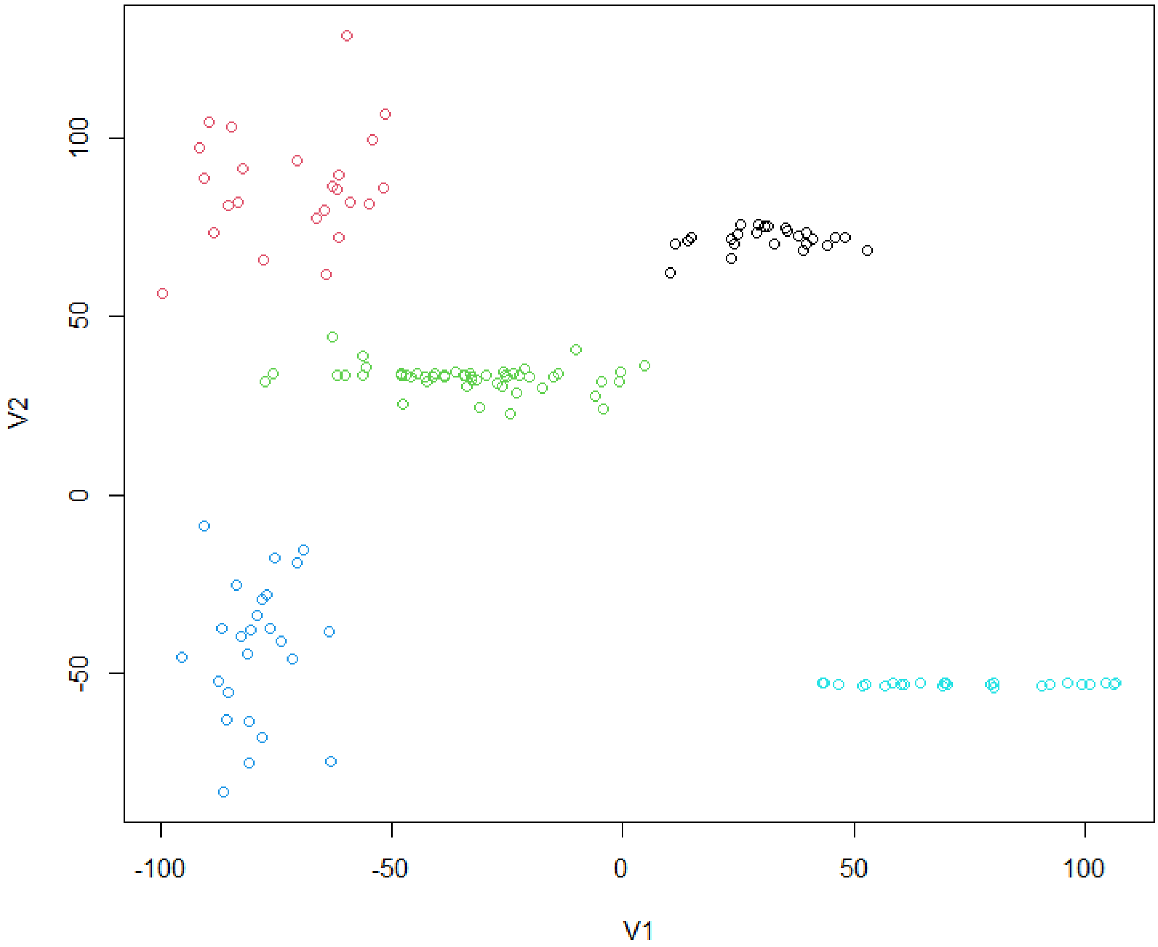
```
Within cluster sum of squares by cluster:
[1] 10670.589 18327.261 10260.915  3436.462 10864.708
 (between_SS / total_SS =  94.1 %)
```

Iteration 2:

```
Within cluster sum of squares by cluster:
[1]  3436.462 10670.589 18327.261 10864.708 10260.915
 (between_SS / total_SS =  94.1 %)
```

```
with(d, plot(V1, V2, col=km$cluster))
```

With the quality measure 94.1% diagram shows expected clusters

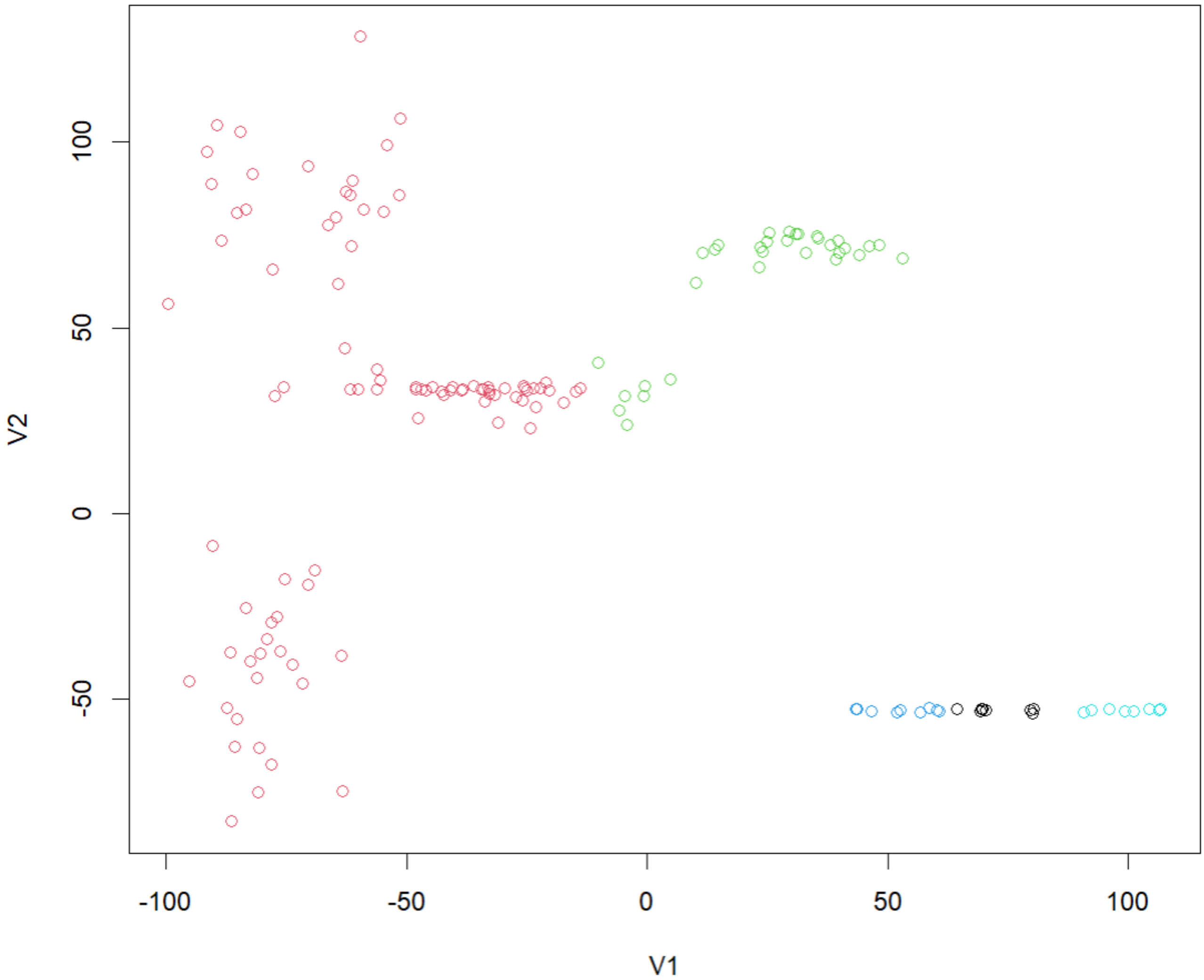


Iteration 3:

```
Within cluster sum of squares by cluster:
[1]     266.0465 274338.3389  18684.9885     385.3310     267.4781
```

```
 (between_SS / total_SS =  67.9 %)
```

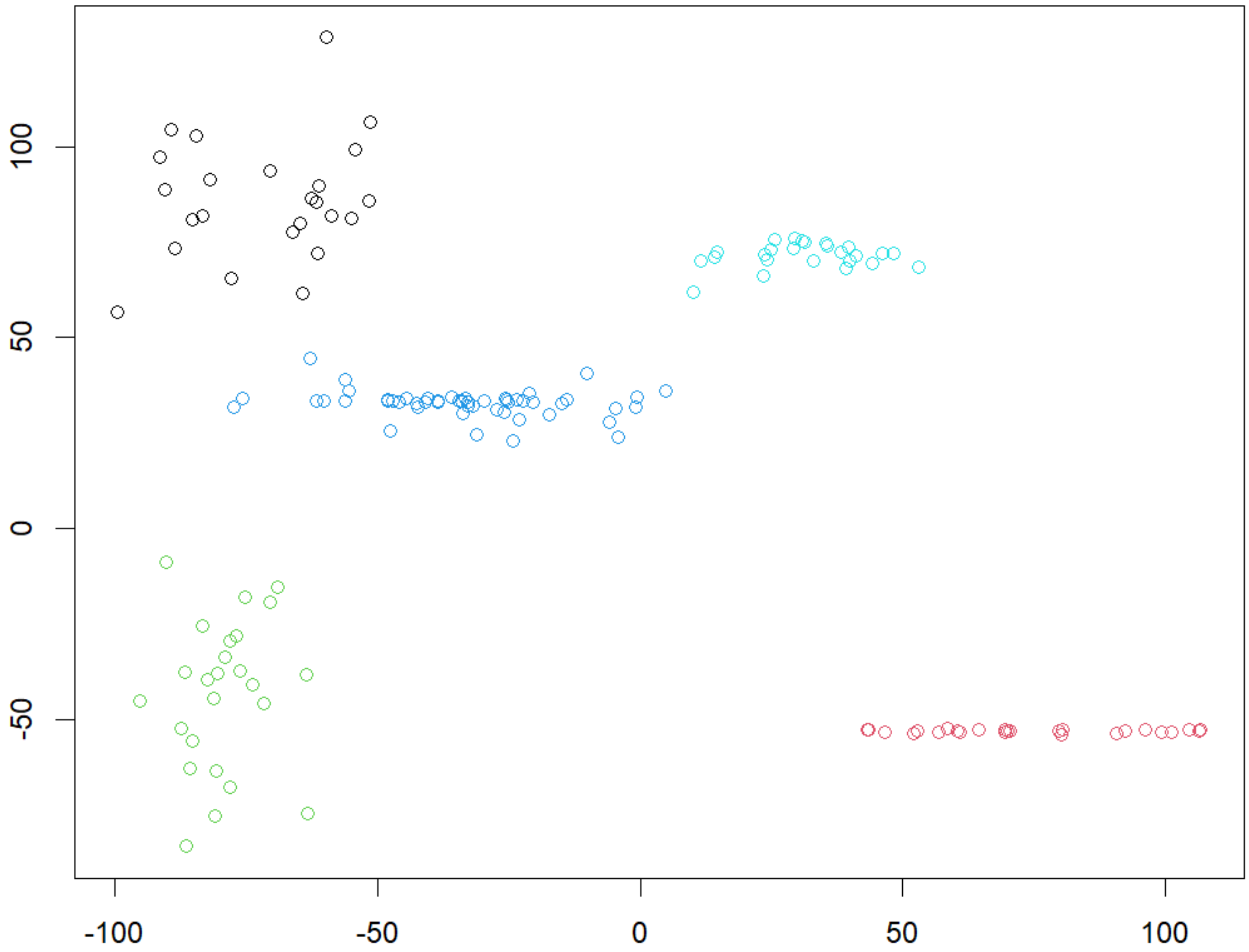
Quality measure is 67.9 which is worse than iteration 1 and 2. The diagram shows that algorithm identified 3 clusters in the right-bottom corner (where supposed to be only one) and remaining data split between remaining two clusters. For the greater number of clusters this might make sense, but we have only 5 so result is not very good.



Iteration 4:

```
Within cluster sum of squares by cluster:
[1] 10670.589 10260.915 10864.708 18327.261  3436.462
 (between_SS / total_SS =  94.1 %)
```

Again, very similar to Iteration 1 and 2



Iteration 5:

```
Within cluster sum of squares by cluster:
[1] 10670.589 10260.915  3436.462 10864.708 18327.261
 (between_SS / total_SS =  94.1 %)
```

And last iteration also shows similar result.

Summary table for all 5 iterations

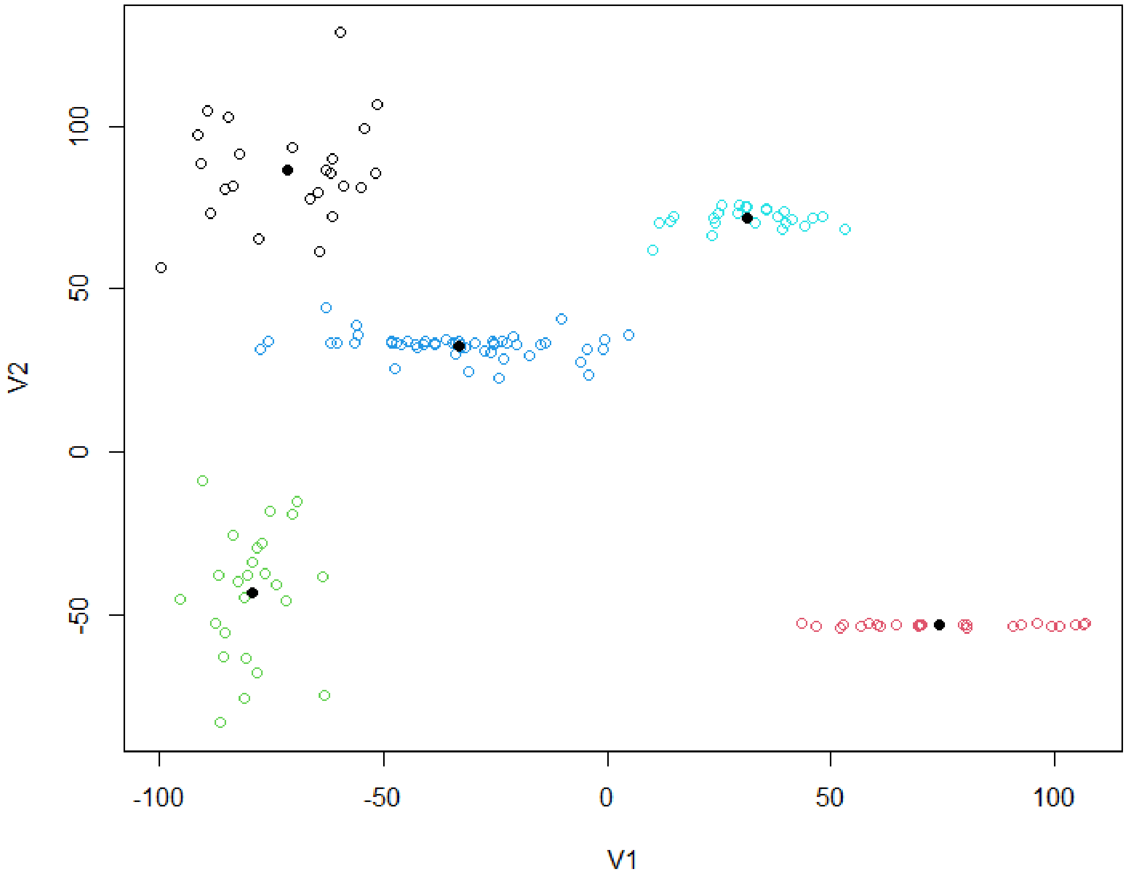
	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
SSB/TSS	94.1%	94.1%	67.9%	94.1%	94.1%

Minimal value: 67.9%

Maximal value: 94.1%

Average value: 88.86%

The solution with centroids shown on the diagram below.



b) Identify the number of clusters using the Elbow Method. Draw the chart that shows how WSS depends on k. Compare the results with the results of Task 1a.

Step 3. Use Elbow Method to identify number of clusters.

Script performs N iterations of k-means algorithm with number of clusters equal current iteration number from 1 to N. Algorithm runs K times in each iteration.

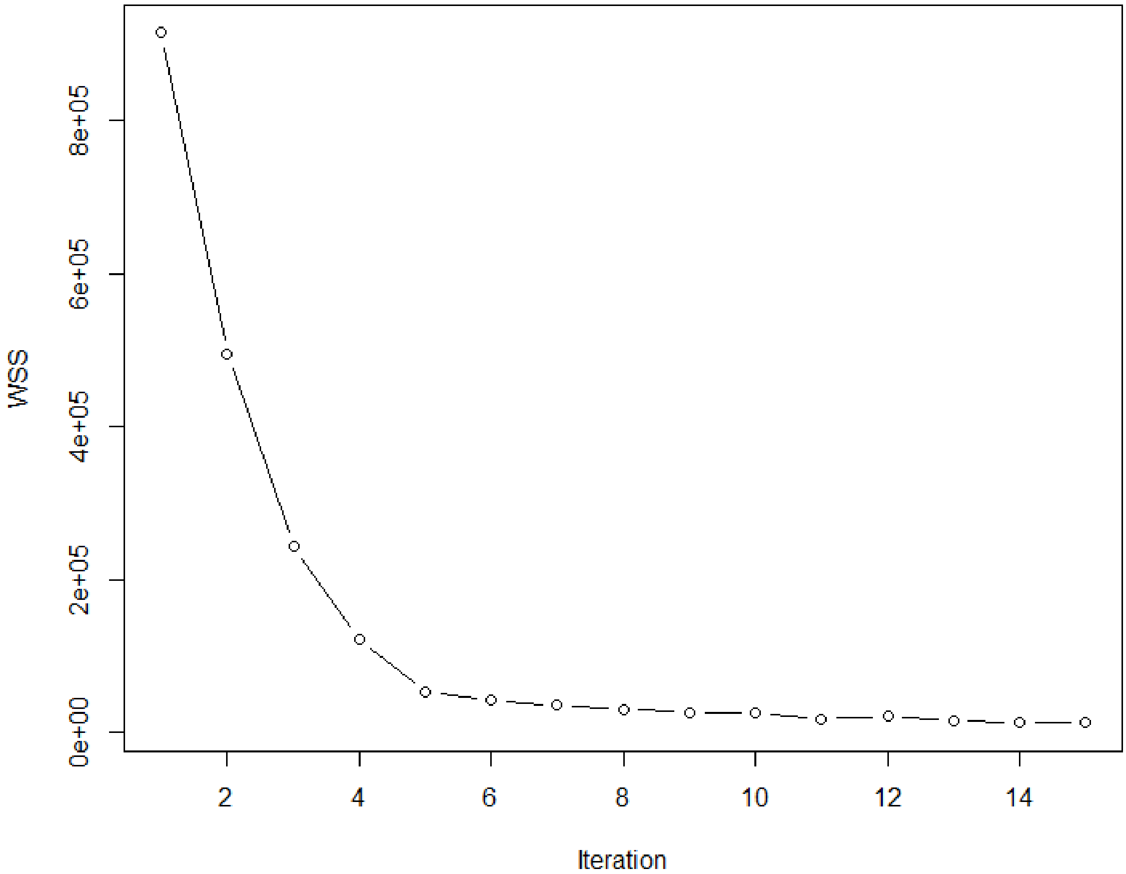
```
N <- 15 # Number of iterations
K <- 5 # Number of algorithm runs in each iteration
results <- list()
for (i in 1:N) {
  km <- kmeans(d, centers = i, nstart = K)
  results[[i]] <- km
}

wss <- sapply(results, function(x) x$tot.withinss)
wss
plot(wss, type = "b", xlab = "Iteration", ylab = "WSS")
```

WSS in each iteration are:

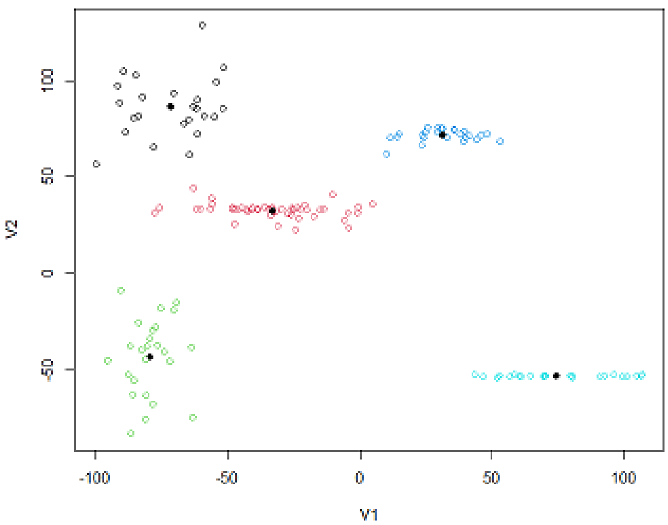
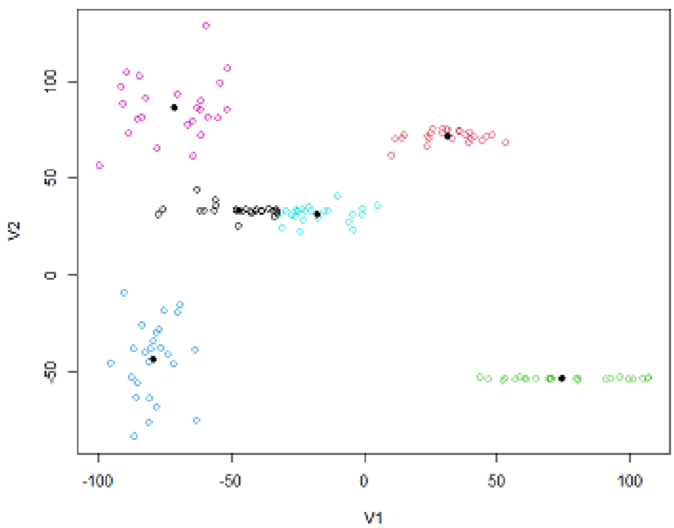
```
[1] 914637.42 [2] 494055.40 [3] 244007.56 [4] 123037.13 [5] 53559.94
[6] 42632.39 [7] 35810.32 [8] 31048.06 [9] 26675.12 [10]  26372.47
[11] 18161.88 [12] 21246.19 [13] 15605.51 [14] 12704.59 [15] 13443.24
```

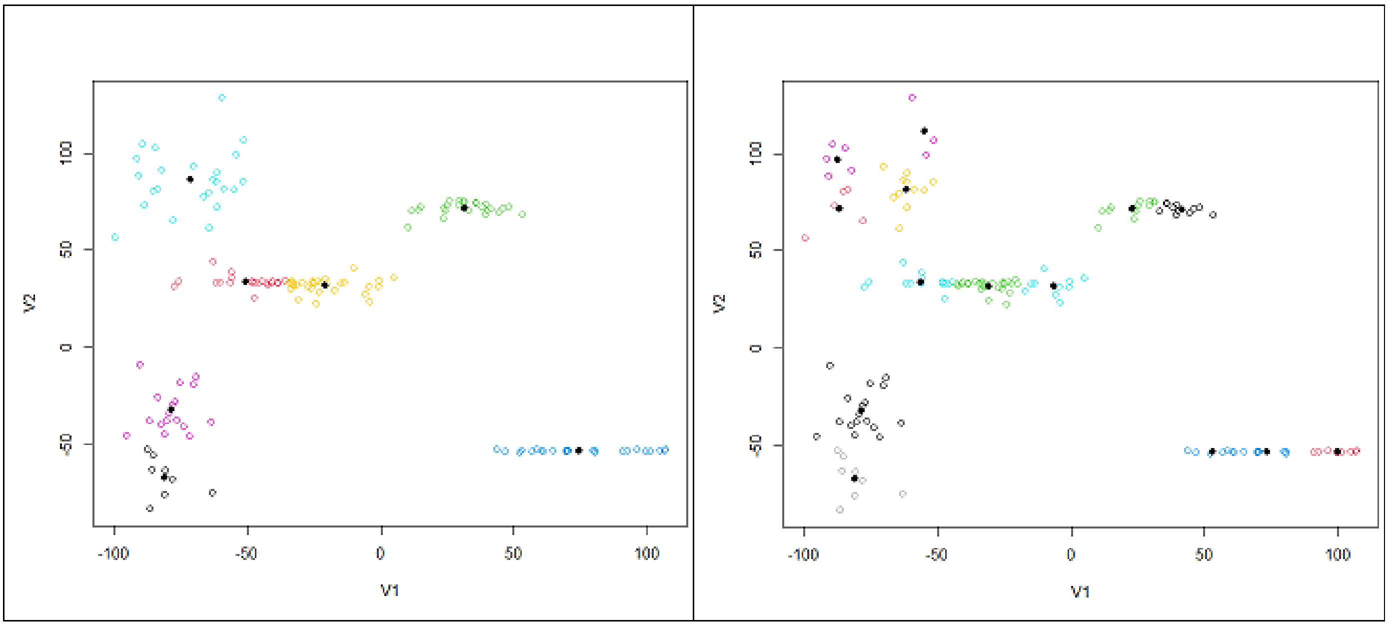
And the chart is on the diagram below. On the diagram, we see that after 5 clusters, the WSS improves very slowly, so our initial assumption about 5 clusters seems correct.



Additionally, we can compare scatter plot for different number of clusters.

```
# plot 5th iteration result
with(d, plot(V1, V2, col=results[[5]]$cluster))
points(results[[5]]$centers, col=1,pch=16,cex=1)
```

5 clusters	6 clusters
	
7 clusters	14 clusters



On the scatter plot representation, we see that 7 clusters might make sense but not very different from 5 clusters. At the same time, 14 cluster definitely looks very redundant.

Task 2:

File "zoo.csv" contains variables describing 7 types of animals: mammal, fish, bird,

invertebrate, insect, amphibian and reptile. Determine clusters in data from "zoo.csv" using the

k-means algorithm. Compare your results with the known classification presented in file

"zoo_full.xlsx" (see the last column "type"). Which of the animals were often misclassified?

Solution:

R script finds 7 clusters in zoo data, binds this information to the dataset and writes new file zoo_clustered.csv for further analysis.

```
d <- read.csv("zoo.csv", header = TRUE, sep = ",")
d

km <- kmeans(d, centers=7, nstart=10)
km
Clustering vector:
  [1] 4 4 5 4 4 4 4 5 5 4 4 1 5 5 7 2 1 4 5 5 1 1 4 1 2 7 7 3 4 3 2 4 3 1 5 4 4
1 5 2 2 1 2 1 4 4 2 4 4 4 4 2 7 6 4 4 1 1 1 1
 [61] 5 5 5 4 4 4 5 4 4 4 4 1 6 5 5 3 5 5 1 1 5 5 5 1 3 7 5 1 2 7 7 7 5 3 4 1 3
2 4 5 1

Within cluster sum of squares by cluster:
[1] 19.100000 12.600000  8.571429 24.451613 44.608696  3.000000 12.375000
 (between_SS / total_SS =  82.2 %)

d_clustered <- cbind(d, cluster=km$cluster)
write.csv(d_clustered, "zoo_clustered.csv")
```

File zoo_clustered.csv imported to Excel and two columns (animal and type) from zoo_full.xlsx added to zoo_clustered data to compare results.

Clusters are:

Cluster 1 – birds, seems correct.

animal	Column1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	cluster	type
chicken	12	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	1	bird
crow	17	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	0	1	bird
dove	21	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	1	bird
duck	22	0	1	1	0	1	1	0	0	1	1	0	0	2	1	0	0	1	bird
flamingo	24	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	1	1	bird
gull	34	0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0	1	bird
hawk	38	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	0	1	bird
kiwi	42	0	1	1	0	0	0	1	0	1	1	0	0	2	1	0	0	1	bird
lark	44	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	1	bird
ostrich	57	0	1	1	0	0	0	0	0	1	1	0	0	2	1	0	1	1	bird
parakeet	58	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0	1	bird
penguin	59	0	1	1	0	0	1	1	0	1	1	0	0	2	1	0	1	1	bird
pheasant	60	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	1	bird
rhea	72	0	1	1	0	0	0	1	0	1	1	0	0	2	1	0	1	1	bird
skimmer	79	0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0	1	bird
skua	80	0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0	1	bird
sparrow	84	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	1	bird
swan	88	0	1	1	0	1	1	0	0	1	1	0	0	2	1	0	1	1	bird
vulture	96	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	1	1	bird
wren	101	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0	1	bird

Cluster 2 – mostly insects, but also invertebrates: “crayfish” and “lobster”

animal	Column1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	cluster	type
crayfish	16	0	0	1	0	0	1	1	0	0	0	0	0	6	0	0	0	2	invertebrate
flea	25	0	0	1	0	0	0	0	0	0	1	0	0	6	0	0	0	2	insect
gnat	31	0	0	1	0	1	0	0	0	0	1	0	0	6	0	0	0	2	insect
honeybee	40	1	0	1	0	1	0	0	0	0	1	1	0	6	0	1	0	2	insect
housefly	41	1	0	1	0	1	0	0	0	0	1	0	0	6	0	0	0	2	insect
ladybird	43	0	0	1	0	1	0	1	0	0	1	0	0	6	0	0	0	2	insect
lobster	47	0	0	1	0	0	1	1	0	0	0	0	0	6	0	0	0	2	invertebrate
moth	52	1	0	1	0	1	0	0	0	0	1	0	0	6	0	0	0	2	insect
termite	89	0	0	1	0	0	0	0	0	0	1	0	0	6	0	0	0	2	insect
wasp	98	1	0	1	0	1	0	0	0	0	1	1	0	6	0	0	0	2	insect

Cluster 3 – mammals. Type seems correct but this cluster inhabitants are very unlikely have many in common.

animal	Column1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	cluster	type
fruitbat	28	1	0	0	1	1	0	0	1	1	1	0	0	2	1	0	0	3	mammal
girl	30	1	0	0	1	0	0	1	1	1	1	0	0	2	0	1	1	3	mammal
gorilla	33	1	0	0	1	0	0	0	1	1	1	0	0	2	0	0	1	3	mammal
sealion	76	1	0	0	1	0	1	1	1	1	1	0	1	2	1	0	1	3	mammal
squirrel	85	1	0	0	1	0	0	0	1	1	1	0	0	2	1	0	0	3	mammal
vampire	94	1	0	0	1	1	0	0	1	1	1	0	0	2	1	0	0	3	mammal
wallaby	97	1	0	0	1	0	0	0	1	1	1	0	0	2	1	0	1	3	mammal

Cluster 4 – all mammals. Nothing strange here.

animal	Column1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	cluster	type
aardvark	1	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	4	mammal
antelope	2	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	4	mammal
bear	4	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	4	mammal
boar	5	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
buffalo	6	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	4	mammal
calf	7	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	4	mammal
cavy	10	1	0	0	1	0	0	0	1	1	1	0	0	4	0	1	0	4	mammal
cheetah	11	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
deer	18	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	4	mammal
elephant	23	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	4	mammal
giraffe	29	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	4	mammal
goat	32	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	4	mammal
hamster	36	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	0	4	mammal
hare	37	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	0	4	mammal
leopard	45	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
lion	46	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
lynx	48	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
mink	49	1	0	0	1	0	1	1	1	1	1	0	0	4	1	0	1	4	mammal
mole	50	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	0	4	mammal
mongoose	51	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
opossum	55	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	0	4	mammal
oryx	56	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	4	mammal
platypus	64	1	0	1	1	0	1	1	0	1	1	0	0	4	1	0	1	4	mammal
polecat	65	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
pony	66	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	4	mammal
puma	68	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
pussycat	69	1	0	0	1	0	0	1	1	1	1	0	0	4	1	1	1	4	mammal
raccoon	70	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal
reindeer	71	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	4	mammal
vole	95	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	0	4	mammal
wolf	99	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	4	mammal

Cluster 5 – Mix of fish, invertebrate, mammal and one reptile - pitviper. Mammals are dolphin, porpoise and seal. Invertebrates are: clam, seawasp, slug, worm.

animal	Column1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	cluster	type
bass	3	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	5	fish
carp	8	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	5	fish
catfish	9	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	5	fish
chub	13	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	5	fish
clam	14	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	5	invertebrate
dogfish	19	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1	5	fish
dolphin	20	0	0	0	1	0	1	1	1	1	1	0	1	0	1	0	1	5	mammal
haddock	35	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0	0	5	fish
herring	39	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	5	fish
pike	61	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1	5	fish
piranha	62	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	5	fish
pitviper	63	0	0	1	0	0	0	1	1	1	1	1	0	0	1	0	0	5	reptile
porpoise	67	0	0	0	1	0	1	1	1	1	1	0	1	0	1	0	1	5	mammal
seahorse	74	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0	0	5	fish
seal	75	1	0	0	1	0	1	1	1	1	1	0	1	0	0	0	1	5	mammal
seasnake	77	0	0	0	0	0	1	1	1	1	0	1	0	0	1	0	0	5	reptile
seawasp	78	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	5	invertebrate
slowworm	81	0	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0	5	reptile
slug	82	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	5	invertebrate
sole	83	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0	0	5	fish
stingray	87	0	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	5	fish
tuna	93	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1	5	fish
worm	100	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	5	invertebrate

Cluster 6 – invertebrates. There must be more, but algorithm selected only two.

animal	Column1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	cluster	type
octopus	54	0	0	1	0	0	1	1	0	0	0	0	0	8	0	0	1	6	invertebrate
scorpion	73	0	0	0	0	0	0	1	0	0	1	1	0	8	1	0	0	6	invertebrate

Cluster 7 – also mix of amphibian, invertebrate and reptile.

animal	Column1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	cluster	type
crab	15	0	0	1	0	0	1	1	0	0	0	0	0	4	0	0	0	7	invertebrate
frog	26	0	0	1	0	0	1	1	1	1	1	0	0	4	0	0	0	7	amphibian
frog	27	0	0	1	0	0	1	1	1	1	1	1	0	4	0	0	0	7	amphibian
newt	53	0	0	1	0	0	1	1	1	1	1	0	0	4	1	0	0	7	amphibian
starfish	86	0	0	1	0	0	1	1	0	0	0	0	0	5	0	0	0	7	invertebrate
toad	90	0	0	1	0	0	1	0	1	1	1	0	0	4	0	0	0	7	amphibian
tortoise	91	0	0	1	0	0	0	0	0	1	1	0	0	4	1	0	1	7	reptile
tuatara	92	0	0	1	0	0	0	1	1	1	1	0	0	4	1	0	0	7	reptile

Conclusions.

1. Sea mammals like dolphin, porpoise and seal mistakenly considered as fish and such an error even humans do.
2. Amphibian, invertebrate and reptiles are difficult to distinguish by the attributes in this dataset.